

CLAIMS:

What is claimed is:

1. A method, in a data processing system, for parsing Eastern Asian language
5 character streams, the method comprising:
 receiving a corpus of word-based parse trees;
 converting the corpus of word-based parse trees into a corpus of character-based
 parse trees; and
 training a parser using the corpus of character-based parse trees.
10
2. The method of claim 1, wherein each word-based parse tree in the corpus of
word-based parse trees includes a word tag for each word in the word-based parse tree.
3. The method of claim 2, wherein converting the corpus of word based parse trees
15 includes assigning a word position tag to each character in the character-based parse tree
based on the word tag for each word in the word-based parse tree.
4. The method of claim 3, wherein the word position tag is one of a beginning tag, a
middle tag, and an end tag.
20
5. The method of claim 1, wherein training the parser includes forming a model.
6. The method of claim 5, further comprising:
 providing the model to a decoder, wherein the decoder parses Eastern Asian
25 language character streams at a character level using the model.

7. The method of claim 6, further comprising:
receiving a test sentence, wherein the test sentence is an Eastern Asian language
character stream; and
parsing the test sentence using the decoder to form one or more character-based
5 parse trees.
8. The method of claim 1, wherein training the parser includes training the parser
using maximum-entropy method.
- 10 9. The method of claim 1, wherein the Eastern Asian language is one of Chinese,
Japanese, and Korean.
10. The method of claim 1, wherein the corpus of word-based parse trees is a Chinese
Treebank.
15
11. An apparatus for parsing Eastern Asian language character streams, the apparatus
comprising:
means for receiving a corpus of word-based parse trees;
means for converting the corpus of word-based parse trees into a corpus of
20 character-based parse trees; and
means for training a parser using the corpus of character-based parse trees.
12. A computer program product, in a computer readable medium, for parsing Eastern
Asian language character streams, the computer program product comprising:
25 instructions for receiving a corpus of word-based parse trees;

instructions for converting the corpus of word-based parse trees into a corpus of character-based parse trees; and

instructions for training a parser using the corpus of character-based parse trees.

5 13. The computer program product of claim 12, wherein each word-based parse tree in the corpus of word-based parse trees includes a word tag for each word in the word-based parse tree.

10 14. The computer program product of claim 13, wherein the instructions for converting the corpus of word based parse trees includes instructions for assigning a word position tag to each character in the character-based parse tree based on the word tag for each word in the word-based parse tree.

15 15. The computer program product of claim 14, wherein the word position tag is one of a beginning tag, a middle tag, and an end tag.

16. The computer program product of claim 12, wherein the instructions for training the parser includes instructions for forming a model.

20 17. The computer program product of claim 16, further comprising:
instructions for providing the model to a decoder, wherein the decoder parses Eastern Asian language character streams at a character level using the model.

25 18. The computer program product of claim 17, further comprising:
instructions for receiving an input sentence, wherein the input sentence is an Eastern Asian language character stream; and

instructions for parsing the input sentence using the decoder to form one or more character-based parse trees.

19. The computer program product of claim 12, wherein the instructions for training
5 the parser includes instructions for training the parser using maximum-entropy method.

20. The computer program product of claim 12, wherein the Eastern Asian language is one of Chinese, Japanese, and Korean.